Report of Year 1 Activities: Solar Loop Mining to Support Studies of the Coronal Heating Problem (AISR03-0077-0139)

University of Louisville

27th August 2005

Abstract

The Coronal Heating Problem is one of the longest standing unsolved mystery in astrophysics. Measurements of the temperature distribution along the loop length can be used to support or eliminate various classes of coronal temperature models. The temperature analysis of coronal loops is a state-of-the-art astronomy. In order to make progress, scientific analysis requires data observed by instruments such as EIT, TRACE, and SXT. The combination of EIT, TRACE, and SXT information provides a powerful data set that will yield unprecedented detail on the plasma parameters of a variety of coronal loop structures. The biggest obstacle to completing this project is putting the data set together. The search for interesting images (with coronal loops) is by far the most time consuming aspect of this project. Currently, this process is performed manually, and is therefore extremely tedious, and hinders the progress of science in this field. We propose an approach based on data mining to quickly sift through massive data sets downloaded from the online NASA solar image databases and automatically discover the rare but interesting images with solar loops, which are essential in studies of the Coronal Heating Problem. The proposed solar loop mining scheme will rely on the following components: (i) Collection and labeling of a sample data set of images coming from both categories (with and without solar loops), (ii) An optimal feature selection strategy that will facilitate the retrieval task, (iii) A classification strategy to classify the transformed image into the correct class, and (iv) Appropriate measures to validate the effectiveness of the loop mining process. This project will be implemented in three main phases that target the image databases collected by two different instruments, EIT aboard the NASA/European Space Agency spacecraft SOHO and NASAs TRACE. We will leave open the possibility of targeting the SXT database on the Japanese Yohkoh spacecraft if time permits. All the results of this project: literature, software, and mined Semantic loop features and class labels (in ASCII and XML formats) on tested portions of the different instrument databases will be made available to the public and other interested researchers via the World Wide Web.

1 Highlights

As part of Phase I of the project, targeting loops that are outside the solar disk in EIT images, we have developed a solar loop retrieval and mining system, that is able to sift through massive data sets downloaded from online solar image databases and automatically discover images containing solar loops. The system includes provision for fine-tuning the recall/precision trade offs according to the user's desires. Most users will insist on high precision, so that they do not have to discard large proportions of the retrieved images with loops, even if this means that they have to sacrifice some recall. We have developed a system that works in two stages. The first stage uses low level image attributes which are fast to compute, and is intended to be tuned to yield high recall values (> 90%). While the second stage classifier adds more costly shape sensitive attributes, but takes as input only the positively classified (recalled) images from the first stage (hence reducing the number of images that have to be processed through the second stage), and can be tuned to yield high precision values (> 90%). The developed system consists of three main subsystems:

- Image Acquisition Subsystem: For retrieving images from the NASA repository and marking the images with solar loops.
- Multi-stage classifier training subsystem. For training classifiers (using low and high level attributes) capable of detecting loops in images.
- Loop Mining Subsystem. This is the production mode loop image retrieval system that uses the trained classifiers on new images directly downloaded from the NASA repository.

Some of the desirable features of the developed system include:

- **Automation and Streamlining the Entire Solar Loop Mining Process:** The solar loop mining cycle has been automated all the way from the image downloading to the marking and label extraction, region of interest extraction, preprocessing, training, testing, and validation. This is enabled mostly thanks to scripting and automating all the phases, including the development of user friendly (click and drag based) interactive or batch style (depending on the user's choice) graphical tools that help download solar images in small or big batches in specific date intervals.
- **Automated and User Friendly Image Acquisition Subsystem:** A tool that allows a user to download solar images in small or big batches from online Web databases, in specific date intervals. The download tool can be used in the learning/training or in the testing/deployment phase. Also, a tool used only in the training phase, that allows a user to mark the loop positions in the images, and saves this information as an addendum to the FITS header. After the marking is done, a set of image blocks, tangent to the sun circle, are automatically extracted from each preprocessed image. Only blocks that are tangent to the sun's circumference are extracted since the scope of the project in Phase I is to detect out of disk solar loops. Furthermore, the optimal block size and block positions

ALGORITHM	TYPE
Decision Stump	Tree Induction
C4.5 (J4.8)	Tree Induction
RepTree	Tree Induction
Conjunctive Rule	Rule Generation
Decision Table	Rule Generation
PART	Rule Generation
JRip	Rule Generation
1-NN	Lazy
3-NN	Lazy
SVM	Function based
ML Perceptron	Function based
Naive Bayes	Probabilistic

Table 1: Tested Algorithms.

are determined automatically to minimize the chance of broken loops (among several blocks). These blocks serve as the elementary objects used in a "learn by example" data mining framework in the next stages.

- **A Multi-Stage Classifier:** The idea of the first stage classifier is to filter out those blocks that are unlikely to contain loops. It is very important to have a classifier with a very low False Negative rate (or equivalently high True Positive rate or *Recall*), since we don't want to lose blocks containing loops in this stage. The goal of this subsystem is to produce computational classifier models, which are able to discriminate image blocks that contain loops from those that do not contain loops. This process is divided in two main stages depending on the features used: (i) Low level features : such as intensity, texture, etc, (ii) High-level semantic structures (solar loops): Shape recognizers, Hough transform, scalable clustering for large data sets (single pass, robust to noise). Because of the high computaional cost of High-level features, we have restricted the first stage classifier to use only low level features, while allowing higher level but costlier features in the second stage, where the stream of images to be sifted through would have at least been abated somewhat. Table 1 presents the different classification algorithms that were used in this work. In order to determine the advantage of applying preprocessing, different experiments were performed: without preprocessing, applying despeckling, and applying both despeckling and Gadient transformation. The latter approach proved to be the most optimal.

- **A Cost-Sensitive Classifier:** In general, a classification technique tries to minimize the global error of the classifier. In our case, this means that we will get a low recall since the negative class (no loop) is bigger than the positive class (the positive samples account for around a 20% of the data set). Therefore, we had to modify a given classifier by biasing it towards producing fewer false negatives,

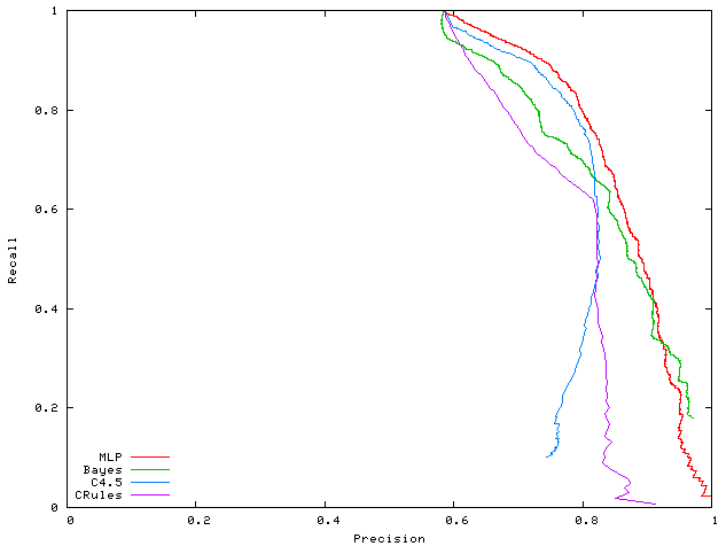
even at the cost of increasing the false positives. We had investigated three main possibilities to do this: (i) Changing the decision threshold for those classifiers that return a vector with the probability of each class, (ii) Modifying the training algorithm to take into account the cost of each type of error, and then giving a higher value to the cost of producing false negatives, (iii) Using an external method (a meta-learner) to make any learning algorithm cost-sensitive.

Figure 1 shows the performance of the top four classification algorithms at the end of stage 2. Since the purpose of the system is to find images that have loops and to identify the area of the image where the loop is located, it seemed natural to evaluate the classifier in terms of the loop-containing regions correctly identified, rather than evaluating individual blocks. For this reason, we have furthermore, refined our classification labels to take into account the regional properties of the blocks in an image so that the final classification will be based on whether any of several adjacent blocks in the same region are classified as loop or no-loop blocks. So far, the tested classifier was tuned to produce a high recall (in terms of blocks) to minimize the chance of missing positive images, and at this setting, we obtained about 98% recall and 68% precision. However, if we change the cost parameter of the classifier, then we can expect the precision to increase without sacrificing much recall. Since some users would rather sacrifice recall for the benefit of precision (in a real life setting, most scientists would rather prefer retrieving half of the images, provided that the greatest majority of them are correctly identified as containing loops), it may also be worthwile to tip the balance towards precision, in which case, we can achieve more than 90% precision, meaning that 9 out of each set of 10 images that are retrieved will be of interest. We are currently still in the process of investigating different approaches to improve our results in the second stage towards a perfect precision rate, namely by adding a local shape analysis procedure.

Finally, we developed a robust statistical estimator for data stream environments (ACRES-STREAMS) yielding a single-pass robust statistical estimator that is free of assumptions about the noise contamination rate and scale value. Robust assumption-free scale estimation allows us to benefit from distribution-independent Chebyshev bounds for outlier and emerging cluster detection. Our rigorous influence function based theoretical analysis concluded that this technique is robust and efficient. Hence, it is expected to extract loop information reliably and in a single pass from a noisy Hough space, which is a well known and challenging task in computer vision.

2 Relevance to NASA

The search for interesting images for coronal temperature analysis (with coronal loops) amounts to searching for a needle in a haystack, and therefore hinders the fast progress of science in this field. The next generation EIT called MAGRITE, scheduled for launch in a few years on NASAs Solar Dynamics Observatory, will require state of the art techniques to sift through the massive wealth of data to support scientific discoveries. The proposed work addresses goals 1 and 2 of the Applied Information Systems Research (AISR), since it includes novel information technology and computational methods that promise to increase productivity of the OSS research and public outreach



(a)

Figure 1: Recall-Precision tradeoff Curves of the top four classification techniques on the second stage data set

endeavors, and would benefit the state-of-practice in space science. It also fosters inter-disciplinary collaboration spanning the space science (Co-I) and computer science (PI) disciplines. Our project addresses objective 4 of the AISR Program, namely increasing science and educational return from the data through advanced knowledge discovery methodologies.

3 Application to NASA Missions and Programs

The Coronal Heating Problem is one of the longest standing unsolved mysteries in astrophysics. Measurements of the temperature distribution along the loop length can be used to support or eliminate various classes of coronal temperature models. The temperature analysis of coronal loops is a state-of-the-art astronomy. In order to make progress, scientific analysis requires data observed by instruments such as EIT, TRACE, and SXT. The combination of EIT, TRACE, and SXT information provides a powerful data set that will yield unprecedented detail on the plasma parameters of a variety of coronal loop structures. The biggest obstacle to completing this project is putting the data set together. The search for interesting images (with coronal loops) is by far the most time consuming aspect of this project. Currently, this process is performed manually, and is therefore extremely tedious, and hinders the progress of science in this field. Our project aims to accelerate and automate the discovery of the rare but interesting images with solar loops.

In addition to the specific problem from Astrophysics, above, research that advances state of the art in solar physics will have a significant impact on society and other scientific fields because of the following reasons: (i) The climate connection: the sun is a source of light and heat for life on Earth. Scientists strive to understand how it works, why it changes, and how these changes influence the Earth, (ii) Space weather: The sun is the source of the solar wind: flow of gases from the sun that streams past the Earth at speeds exceeding a million miles per hour. Disturbances in the solar wind shake the Earths magnetic field. Space weather can change the orbits of satellites and shorten mission lifetimes. Excess radiation can physically damage satellites and poses a threat to astronauts, in addition to power surges and outages on Earth, and hence needs to be predicted. (iii) The sun as a physical laboratory: the sun produces its energy by nuclear fusion, a process that scientists have strived for decades to reproduce by involving hot plasmas in strong magnetic fields. Much of solar astronomy involves observing and understanding plasmas under similar conditions.

4 Tracking

All the results of this project: literature, software, and outputs (labels) of the developed classification methods on tested portions of the different instrument databases will be made available to the public and other interested researchers via the World Wide Web. Outputs of our automated retrieval process (on new test data) will be saved in both ASCII column format and XML format to facilitate data interchange with and between different research groups. The XML schema will include derived Semantic features (Estimated number of loops and their confidence) and the assigned class labels.

Tracking the usage of our products is easily accomplished by monitoring the access statistics on the projects website, a feature that is already in use, as well as searching for citations on the Web.

5 Software and Publications

Our software is currently on the project collaboration platform website:

"http://webmining.spd.louisville.edu/twiki/bin/view/SOLARLoops/".

Since we are in the first year of this project, we intend to publish our developed systems and results to a larger audience, once it is completely stable, fully documented, and bug free. Again, because of the early stages of this project, our publications are mostly in preparation and will be submitted, as follows: a long paper to be submitted to the "*SigKDD Explorations journal - Special Issue on "Success Stories in Data Mining*", a shorter more focused paper about the developed approach and system to the *KDD Applications/Industrial track*, and a more theoretically and algorithimically inclined paper on the developed evolving stream clustering approach, ACRES-Streams to the *SIAM Data Mining Conference*.

6 Upcoming Plans

- Scaling the Hough transformation procedure by incorporating constraints to prune the space of possible parameters in the early stages of computation
- Integration of the Hough space analysis with single-pass robust clustering (ACRES-Streams) to automate the semantic loop feature extraction
- Investigating alternative loop detection methods, such as by seeking clusters of elliptically shaped arcs.
- Continuing the collection and labeling of relevant images (our collaborators in the Solar Physics lab at the University of Memphis), notably in several categories (low, medium, high level of solar activity) depending on the solar cycle (minimum to maximum)
- Applying clustering (unsupervised categorization) on the varying cycle data (different levels of solar activity) to create homogenous groups of images
- Developing context-sensitive loop retrieval mechanisms that automatically adapt to the category of each input image
- Developing classification models using the collected data that take as input a combination of the extracted loop features together with intrinsic image features and extrinsic contextual features (from the FITS header fields), such as date to infer solar cycle information.